RSI-72-03

PATTERN RECOGNITION SYSTEMS AND PROCEDURES

Remote Sensing Institute
South Dakota State University
Brookings, South Dakota 57006

February, 1972

RSI-72-03

NASA ANNUAL REPORT

PATTERN RECOGNITION SYSTEMS AND PROCEDURES

BY

GERALD D. NELSON AND DAVID V. SERREYN

TO

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

GRANT NUMBER

NGL 42-003-007

REMOTE SENSING INSTITUTE
SOUTH DAKOTA STATE UNIVERSITY
BROOKINGS, SOUTH DAKOTA

in cooperation with

DEPARTMENT OF ELECTRICAL ENGINEERING
SOUTH DAKOTA STATE UNIVERSITY
BROOKINGS, SOUTH DAKOTA

FEBRUARY, 1972

ABSTRACT

The objectives of the pattern recognition tasks are to develop (a) a man-machine interactive data processing system and (b) procedures to determine effective features as a function of time for crops and soils.

The Signal Analysis and Dissemination Equipment, SADE, is being developed as a man-machine interactive data processing system. SADE will provide imagery and multi-channel analog tape inputs for digitation and a color display of the data. SADE is an essential tool to aid in the investigation to determine useful features as a function of time for crops and soils.

Four related studies are: (1) reliability of the multivariate Gaussian assumption, (2) usefulness of transforming features with regard to the classifier probability of error, (3) advantage of selecting quantizer parameters to minimize the classifier probability of error and (4) advantage of using contextual data.

The initial objective of developing an interactive system will be almost complete at the end of April with the installation of SADE (Signal Analysis and Dissemination Equipment). Classification programs already developed such as K CLASS, Bayes and contextual data will be utilized with SADE. Boundary detection algorithms which are being worked on will also be utilized. In order to speed up processing, a variable quantizer with parameters which are specified by the operator, has also been implemented.

The study of transformation of variables (features), especially those experimental studies which can be completed with the SADE system, will be done. The multivariate Gaussian assumption will either be justified and/or nonparametric techniques will be applied to imagery data. The secondary objective of determining effective features for crops and soils will not be achieved until the SADE system is operational.

INTRODUCTION

The objectives of the pattern recognition tasks are to develop (a) a man-machine interactive data processing system and (b) procedures to determine effective features as a function of time for crops and soils.

This paper reports on the progress made toward achieving these objectives. The specific projects done and currently in progress are considered necessary to aid in the specification of data processing and classification techniques.

To assume that the data or features are multivariate Gaussian is sometimes unreliable. Therefore, in computer simulation studies when pseudo-random number generators are used to generate the feature vectors with specified statistical parameters, a criteria to judge the probability of the data being from the specified probability density function is required. The Kolmogorov-Smirnov test is used to determine if the pseudo-random number generator produced Gaussian data.

Another problem area is to determine the advantages and disadvantages of transforming the measurements. Experimental work by Nalepka (1) has shown the advantage of the ratio technique.

In this paper the ratio transformation is considered. However, the effect of noise is implied to be additive. The case of multiplicative noise should also be pursued.

The selection of quantizer parameters for a two-class probability of error in classification experiments. A study to evaluate the conditions necessary to provide this decreased probability of error is in progress. The purpose and method of the study is discussed.

The man-machine interactive data processing system is referred to at the Remote Sensing Institute as the Signal Analysis and Dissemination Equipment, SADE. This system is described and plans for its use indicated.

The work on transformations of Gaussian variates (3) was done by John E. Boyd, previously a graduate student in the Electrical Engineering Department at South Dakota State University.

The remainder of the reported work was done at the Remote Sensing Institute at South Dakota State University in Brookings, South Dakota. The work was performed under grant number NGL 42-003-007 which is supported by the Office of University Affairs and the Earth Observations Office of NASA.

MULTIVARIATE GAUSSIAN ASSUMPTION

A very common assumption made by investigators is that the features used to represent the pattern classes are multivariate Gaussian. To validate the Gaussian assumption is not a trivial task. Papoulis (4) discusses the bivariate Gaussian case. If the joint probability density function is bivariate Gaussian it is also true that the marginal probability density functions are also Gaussian. However, if the marginal probability density functions are Gaussian the joint probability density function is not necessarily bivariate Gaussian.

To study by computer simulation methods the effectiveness of contextual data requires the use of a pseudo-random number generator. The generation of Gaussian random variates can be conveniently done by the use of the IBM subroutines GAUSS and RANDU. RANDU generates a uniform random number between zero and one by a power residue method. GAUSS requires twelve uniform random numbers to generate one Gaussian random variate. The central limit theorem is used with the number of uniform random numbers set at twelve instead of approaching infinity in order to make the procedure feasible.

Another method of generating Gaussian random variates is known as the Muller method (5). The equations which relate T, Y, U and V are

$$T = \sqrt{-2\ln U} \quad \cos(2\pi V)$$

$$Y = \sqrt{-2\ln U} \quad \sin(2\pi V)$$

The validity of the computer simulation study depends on the probability density function (pdf) of the data produced by the random number generators. To test the pdf's the Kolmogorov-Smirnov test

(K-S test) (6) was used on RANDU, GAUSS and the Muller produced pseudo-random numbers. The results of the K-S test as a function of the number of samples are presented in Table I.

The results of the K-S test indicate that except in several cases the pseudo-random numbers have less than 90 percent chance of being from either uniform or Gaussian pdf's.

GAUSS uses the output of RANDU and the Muller method uses the output of RANDU twice, with different seeds as listed in Table I. The means and variances for RANDU are very good estimates of the specified population means and variances. The results of the K-S test indicate that before the computer simulation experiment is performed the pseudo-random number generator should be evaluated by this test.

TRANSFORMATIONS OF GAUSSIAN VARIATES

In this two-class problem the features selected are denoted X_1 and X_2. For class one these features are uncorrelated, and each is Gaussianly distributed with mean zero and variance one. For class two these features are correlated, and each is Gaussianly distributed with mean zero and variance one. Therefore class one and two are overlapping bivariate Gaussian probability density functions. The contours of these two overlapping probability density functions are shown in Figure 1.

To classify the data based on the features X_1 and X_2 the Bayes classifier was derived and the decision boundaries determined. Figure 2 represents the three sigma contours of the class one and two bivariate Gaussian probability density functions. The decision boundaries are shown as hyperbolas. The alpha and beta errors are .216 and .053 respectively with a total error of 0.1345.

It can be shown (4) that Z, the ratio of two Gaussian variates, $X_1/X_2=Z$ has an univariate Cauchy probability density function. The pdf of X_1 and X_2 is

$$f(X_1,X_2)= \frac{1}{2\pi\,\sigma_1\sigma_2\sqrt{1-r^2}} \exp\left[\frac{-1}{2(1-r^2)}\left(\frac{X_1^2}{\sigma_1^2} - \frac{2rX_1X_2}{\sigma_1\sigma_2} + \frac{X_2^2}{\sigma_2^2}\right)\right]$$

For class one the correlation coefficient r is zero and for class two is .98.

The Cauchy pdf of Z is

$$f(Z) = \frac{\sqrt{1-r^2}\;\sigma_1\sigma_2}{\pi[\sigma_2^2(Z-r\frac{\sigma_1}{\sigma_2})^2 + \sigma_1^2(1-r^2)]}$$

The classification process is now based on the feature Z which is the ratio X_1/X_2. The probability of error is of prime interest. The Bayes classifier for Cauchy data was derived and the results determined. The alpha and beta errors are .347 and .130, respectively. See Figure 3. The total probability of error is .2385. Therefore, under the assumption of this problem it is obvious that taking the ratio is not useful to decrease the probability of error. In fact the probability of error has increased 10.4 percent from 13.45 percent to 23.85 percent.

QUANTIZATION

The effects of using quantized data on classification error was investigated for a two-class problem involving a single feature or attribute. The results of this study are summarized here and given in a Technical Interim Report (7).

A quantizer is best described by its transfer characteristic. The transfer characteristics for an even and odd equi-interval quantizer are shown in Figures 4 and 5 respectively. The input to the quantizer consists of data whose distribution is assumed to be normal. The effect of the quantizer is to assign a specific value to any of the data that falls in a given range. The value for r, the quantization interval, is constant for a given quantizer but varies with the number of levels.

The output of the quantizer consists of NQ values where NQ is the number of quantize levels. Each of these NQ values is weighted by the area under the normal curve within the input range as given by the transfer characteristic. For the multi-sample problem, the pdf (probability density function) is multinomial. This is due to the fact that the output consists of NQ levels. If only two levels are present, then the sampled data is distributed binomially.

For the one sample case, the pdf for the data is the output pdf of the quantizer. The probability of error is given by

$$P(E) = q_1\alpha + q_2\beta$$

where q_1 and q_2 are probabilities of occurrence for each of the two classes. At the output of the quantizer, the probability of error is a summation given by

$$P_q(E) = (q_1/2)\sum_{k=N+1}^{NQ} [\mathrm{erf}\{u_{k+1}+s)/\sqrt{2\sigma}\} - \mathrm{erf}\{(u_k+s)/\sqrt{2\sigma}\}]$$

$$+ (q_2/2)\sum_{k=1}^{N} [\mathrm{erf}\{(u_{k+1}+s-\mu)/\sqrt{2\sigma}\} - \mathrm{erf}\{(u_k+s-\mu)/\sqrt{2\sigma}\}].$$

for the normal case. This reduces to

$$P_q(E) = (q_1/2)[1-\mathrm{erf}\{(u_{N+1}+s)/\sqrt{2\sigma}\}]$$

$$+ (q_2/2)\ 1+\mathrm{erf}\{(u_{N+1}+s-\mu)/\sqrt{2\sigma}\}].$$

In the above equations s is a shift factor, u_{N+1} is the location of the $(N+1)^{th}$ input location. The decision boundary is at the N^{th} impulse. The probability of error is graphed in Figure 6 for various values of N. The entire graph for each value of N is not given. The location of equal probability of error are drawn in. The X locates the minimums.

To find the location of the minimums involves taking the derivative setting to zero and solving for the variable in question, namely r. Solving for r yields

$$r = \{(2\sigma^2/\mu)\ln(q_1/q_2) + \mu-2s\}/(2N-NQ).$$

This value of r gives the minimum probability of error. In fact, this value makes the probability of error equal to that of the continuous case. In Figure 7, the probability of error is graph for two values of r. One case is the derived while the other involves an r that minimizes the mean-square-error between input and output derived by Max (8).

As the number of samples increases before a decision is made, the probability of error decreases in the continuous case. The same is true of error for the multi-sample case is shown in Figure 8 for a two and three level quantizer as well as continuous. Ones does get more probability of error in the quantized case. However, the

differences in error may be acceptable if one is concerned with the accuracy of the measurement and the cost of obtaining such accuracy.

CONTEXTUAL DATA

The data that occurs adjacent to the cell of the image to be classified provides additional information to be combined with the data of the cell. This additional data is known as the contextual data, and the information added is contextual information. The contextual data is added to the decision rule of the classifier by the following product of sums of probability density functions (2).

$$\prod_{i=1}^{4} \sum_{\theta_{k_i}} p(X_{k_i}|\theta_{k_i})\, p(\theta_{k_i}|\theta_k)$$

where

$p(X_{k_i}|\theta_{k_i})$ is the conditional pdf of the measurement vector X_{k_i} given the cells identity, θ_{k_i}.

$p(\theta_{k_i}|\theta_k)$ is the conditional pdf of the occurrence of the θ_{k_i} class given the k^{th} cell identity, and

k_i's are the nearest neighbors to the k^{th} cell.

The complete decision rule is to

$$\text{Minimize} \sum_{\theta_k} L(\theta_k,a)\; p(X_k|\theta_k)G(\theta_k) \prod_{i=1}^{4} \sum_{\theta_{k_i}} p(X_{k_i}|\theta_{k_i})p(\theta_{k_i}|\theta_k)$$

The terms are:

θ_k is the k^{th} class.

X_k is the optical density measurement of the k^{th} class.

a is $(a_1, a_2, - - - a_n)$.

a_k is the decision that the k^{th} class is present.

$L(\theta_k,a)$ is the loss associated with making a decision.

The assumptions are that contextual relationships between non-adjacent cells are negligible which can be stated as

$$p(X_b|\theta_c,X_i,\theta_i) = p(X_b|X_i,\theta_i)$$

for all i and cells b and c are nonadjacent. The appearance x_k of a class θ_k is a function only of θ_k, and if θ_k is known neither the nature nor the appearance of any other class provides additional information about X_k. This can be stated as

$$p(X_k|\theta_k,X_i,\theta_i) = p(X_k|\theta_k).$$

For a special case of a temporal signal on line scan data this contextual rule can be more easily interpreted and the effect of the contextual data illustrated if a two-class problem is discussed. The next equation specifies the mathematical operations necessary to make a decision.

$$\frac{p(X_k|\theta_1)}{p(X_k|\theta_2)} > KM \quad \text{decide } \theta_1$$

$$\leq KM \quad \text{decide } \theta_2$$

where $$K = \frac{G(\theta_2)L(\theta_2,a)}{G(\theta_1)L(\theta_2,a)},$$

and

$$M = \frac{[p(X_{k_1}|\theta_1)p_{12}+p(X_{k_1}|\theta_2)p_{22}][p(X_{k_2}|\theta_1)p_{12}+p(X_{k_2}|\theta_2)p_{22}]}{[p(X_{k_1}|\theta_1)p_{11}+p(X_{k_1}|\theta_2)p_{21}][p(X_{k_2}|\theta_1)p_{11}+p(X_{k_2}|\theta_2)p_{21}]}$$

The notation p_{ij} is used for the a priori probabilities $p(\theta_i|\theta_j)$.

If contextual data is not used the decision rule reduces to

$$\frac{p(X_k|\theta_1)}{p(X_k|\theta_2)} > K$$

or M is assumed to be unity.

To determine the value of M only the measurements in the nearest neighbor cells and the a priori probabilities of occurrence of each class as represented by p_{ij} for all i and j are required. The measurement in cell k provides the needed data to evaluate the left hand side of the decision rule. Since the right hand side, the product of K and M varies one can think of contextual data as providing a decision boundary which varies as a function of the measured data.

A computer simulation is nearly completed which will provide a comparison of the probability of error if (a) contextual data are not used and (b) contextual data are used, as well as these comparisons for more than one measurement per cell.

The similarity of this simulation to data to be extracted from imagery should be noted. The cells correspond to data windows whose width will vary according to field size. The use of the nearest neighbor cells can be easily extended to be used by obtaining the data from the preceeding and succeeding scan lines in the respective cells.

The necessity of having line scan data and a procedure to edit the ground truth data into the system so that it can be used effectively is imperative.

To use this technique on imagery, requires knowledge of at least the soils type and/or crop type for all fields within the training set.

Only the two class case has been discussed, but generalization to more classes has been done (2).

SIGNAL ANALYSIS AND DISSEMINATION EQUIPMENT

INTRODUCTION

This section discusses the Remote Sensing Institute's Signal

Analysis and Dissemination Equipment (SADE).

This equipment was designed as a state-of-the-art data analysis system for a medium cost with highly flexible modular design. The maximum resolution of any individual module is not as great as may be achieved at a higher cost. However, the integration of each segment with a medium high resolution has produced a system of outstanding capabilities. The SADE system includes the following features:

high resolution, high quality digitization of black and white and color film transparencies,

35mm, 70mm, 9½ inch single frame or roll films may be accommodated,

multiple frames may be registered with respect to one another conveniently and accurately,

registration of images can be accomplished off-line without interaction with the computer,

analog magnetic tape data can be digitized for computer analysis from one to six channels,

a refresh memory is provided for the display of processed data on the video monitor and for multiple frame registration,

refresh memory is expandable for storage of high frame resolution imaged data and larger display formats,

memory format is Pax II picture processing language (9) oriented for processing within the computer,

processed digital data may be converted to hard copy using a line printer,

control of the system's components in communication with the computer is provided via teletype and system control panel.

SYSTEM CONFIGURATION

The system is configured in five major units:

1. The Spatial Data camera, level slicing, and color display monitor;

2. The image digitizer utilizing an image dissector tube;

3. The data control and conversion unit which contains elements for high speed memory, analog tape conversion and control functions,

4. The Lockheed 417 seven track analog tape recorder, and

5. The Daedalus Film Printer.

SYSTEM FUNCTIONS

The SADE system will provide the following system functions:

Visual display of digitized film data,

Digitized image data transmitted to the computer,

Analog tape data digitized and transmitted to the computer,

Processed image data transmitted to the display monitor through the refresh memory for visual interpretation,

Processed digital data to the Daedalus Printer for 70mm film output,

Registration of images via the digitizer and display monitor, and

System control.

MODES OF OPERATION

The SADE system can be operated in the image data and analog tape data digitization modes. The image can be digitized and displayed on the color monitor, and the digitized data transmitted to the computer memory. A second image can be digitized and through mechanical manipulation be registered by observation of the color display which displays both the first and second images simultaneously through use of the raster interlaced system. After registration the digitized data which represents the second image is transmitted to the computer memory. Likewise, maps may be made, digitized and stored in the computer memory and the image overlayed on the map on the display.

The image can be quantized according to the optimal quantizer derived in this paper and the resulting color encoded image displayed on the monitor. This is one method of classifying the image according to optical density which has been shown to be useful by Frazee (10).

Classification of the data if multiple features are used will be done by K-class, Bayes or contextual data classifier algorithms. The classification results will be color encoded and displayed on the monitor.

The digitized image can also be transmitted to the Daedalus printer for 70mm hard copy.

The analog tape digitization mode of operation provides the capability to digitize from one to six channels of multi-spectral data stored on analog magnetic tape. The data can be displayed one channel at a time on the display. The data can be digitized 6 channels at a time and transmitted to the computer for storage in digital form. Possible inputs to the analog tape conversion unit includes thermal scanner data, multispectral scanner data up to 6 channels and weather satellite image data.

ANTICIPATED USES OF THE SADE SYSTEM

The SADE system was designed as a flexible research tool for data processing and data classification which can be expanded. The SADE system can also be expanded to provide information dissemination capabilities. These information dissemination capabilities include television broadcasts, special systems and output copy. The dissemination system chosen will depend upon the specific application and output form desired by the user. Although the SADE system is designed as a research tool, it is also anticipated that analysis completed in the research phases will be implemented in an operational system for individual users. Thus, a state agency interested in surface water could use, on an operational basis, information from the ERTS-A satellite in their daily decision-making processes.

CONCLUDING REMARKS

The objectives of the pattern recognition tasks are to develop (a) a man-machine interactive data processing system and (b) procedures to determine effective features as a function of time for crops and soils.

The initial objective of developing an interactive system will be almost complete at the end of April with the installation of SADE (Signal Analysis and Dissemination Equipment). Classification programs already developed such as K CLASS, Bayes and contextual data will be utilized with SADE. Boundary detection algorithms which are being worked on will also be utilized. In order to speed up processing, a variable quantizer with parameters which are specified by the operator, has also been implemented.

The study of transformation of variables (features), especially those experimental studies which can be completed with the SADE system, will be done. The multivariate Gaussian assumption will either be justified and/or nonparametric techniques will be applied to imagery data. The secondary objective of determining effective features for crops and soils will not be achieved until the SADE system is operational.

TABLE I K-S Test Results

PROBABILITY THAT DATA HAS SPECIFIED PDF

Number of Samples	RANDU		GAUSS	MULLER
seed	98765	12345	12345	98765 and 12345
100	35.7	43.3	11.1	64.6
200	46.4	88.4	50.0	92.5
300	81.9	98.8	47.1	88.7
400	88.4	77.3	44.9	77.2
500	99.6	28.0	48.6	90.7
600	15.7	10.4		
700	10.3	24.8		
800	6.0	36.1		
900	15.8	29.7		

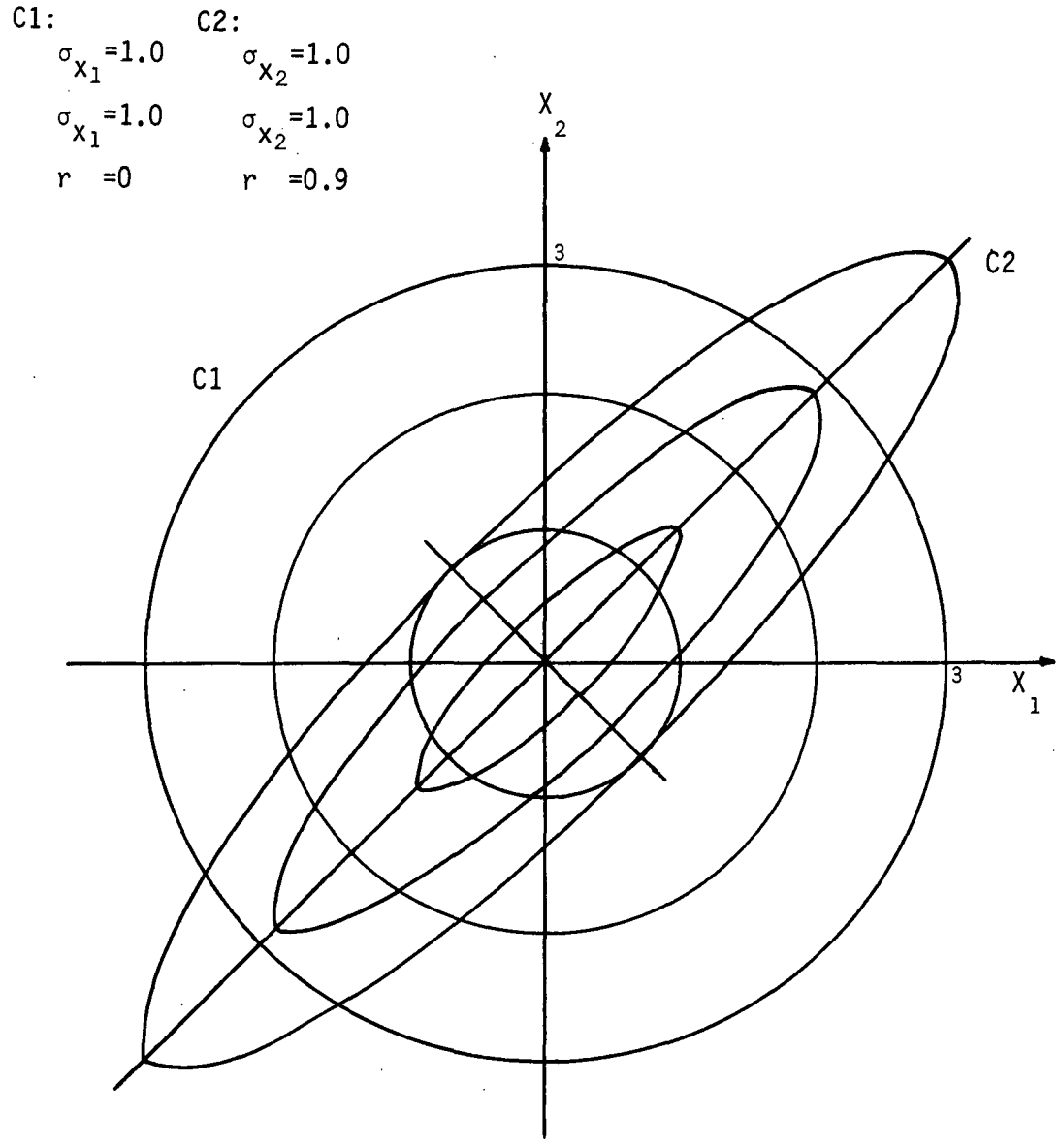

Figure 1.- Contours of two overlapping bivariate Gaussian probability density functions.

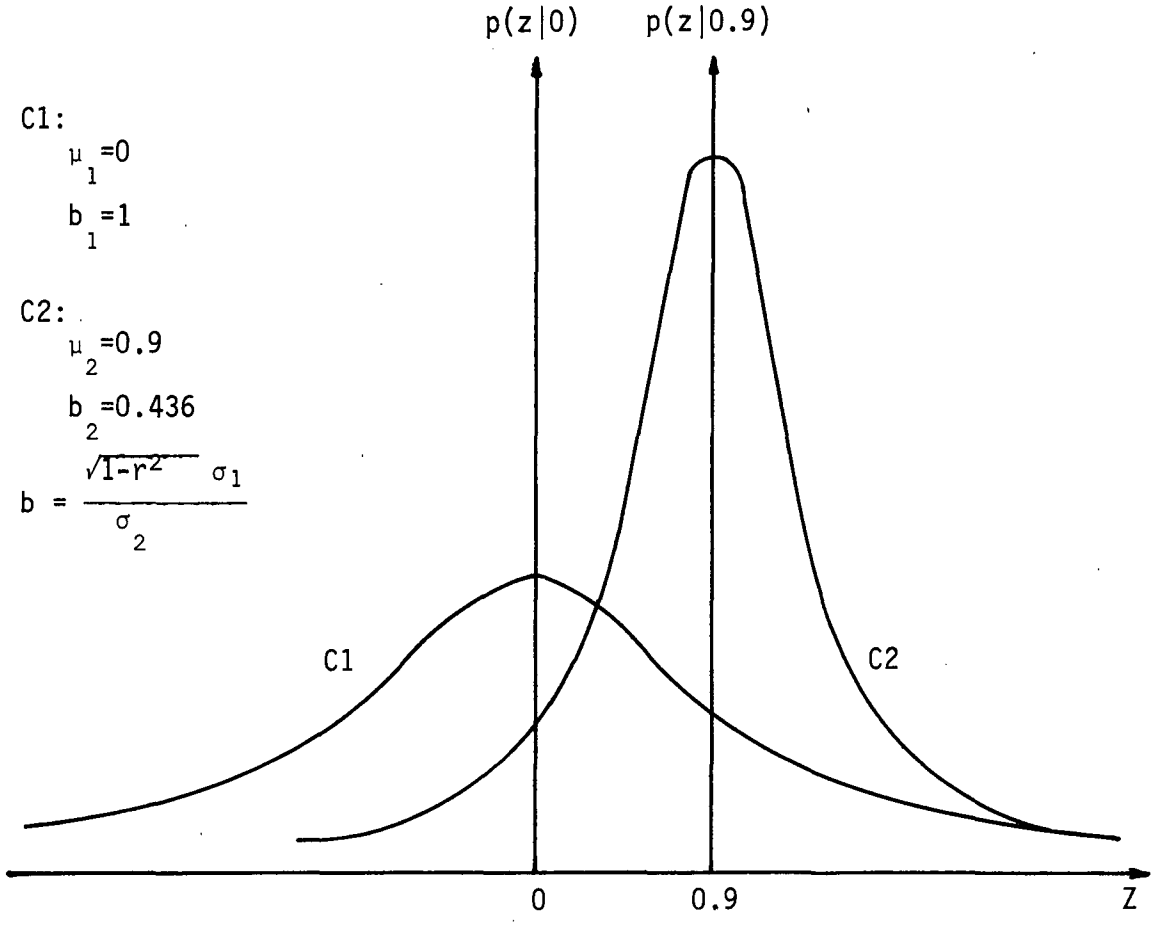

Figure 2.- Two overlapping univariate Cauchy probability density functions.

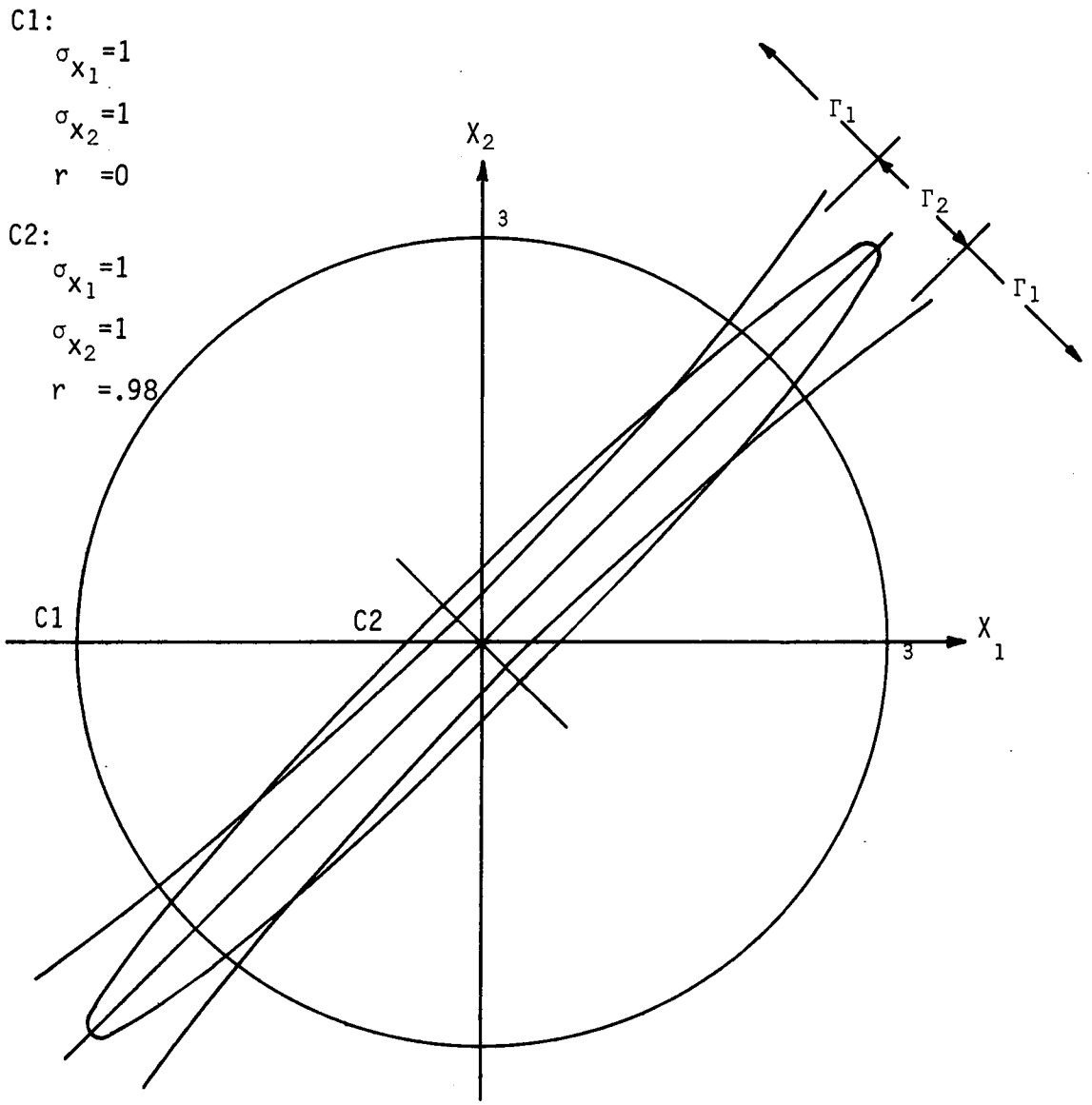

Figure 3.- Decision boundary for Bayes' classification of two bivariate Gaussian distributions, r=0.98.

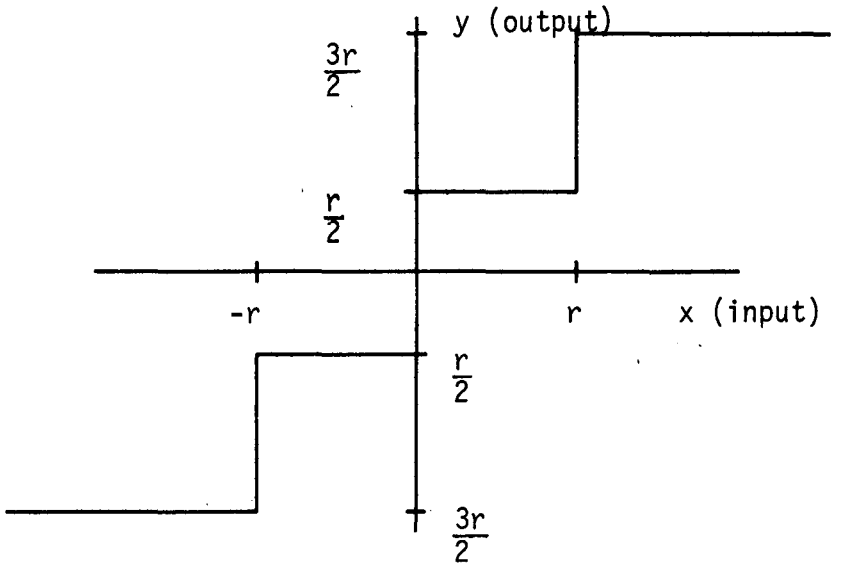

Figure 4.- Transfer characteristic for equiinterval quantizer with an even number of levels (four).

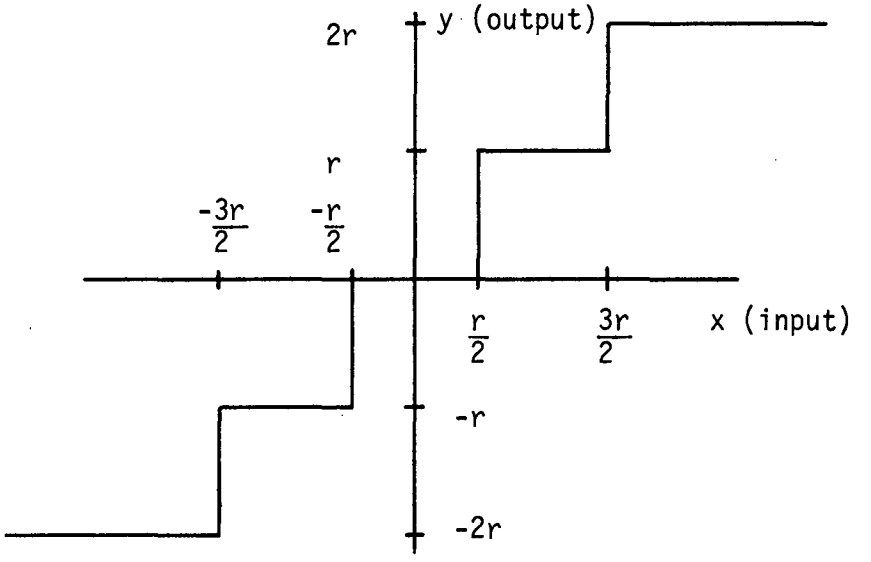

Figure 5.- Transfer characteristic for equiinterval quantizer with an odd number of levels (five).

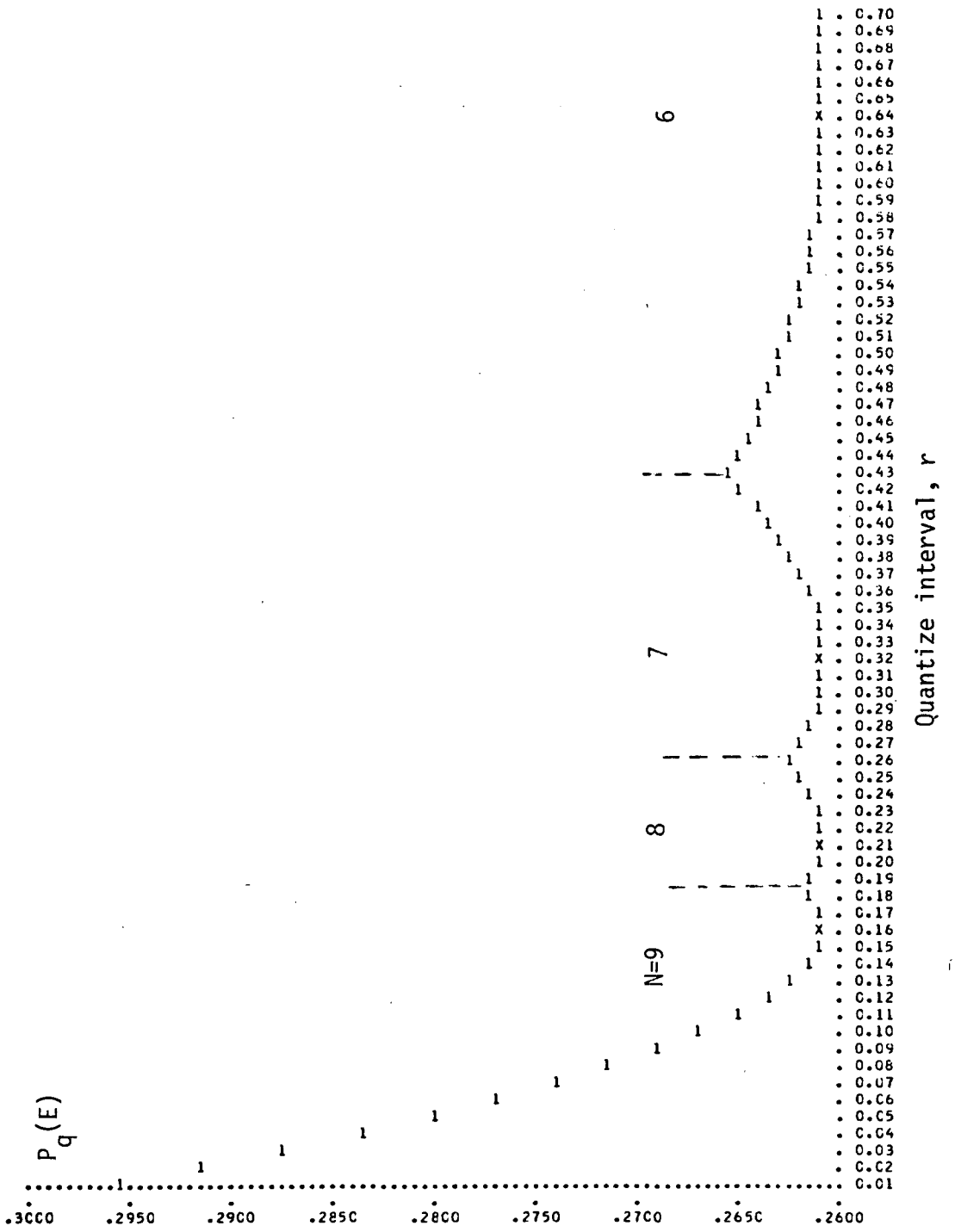

Figure 6.- Probability of error for different decision boundaries for varying quantization interval with a ten level quantizer.

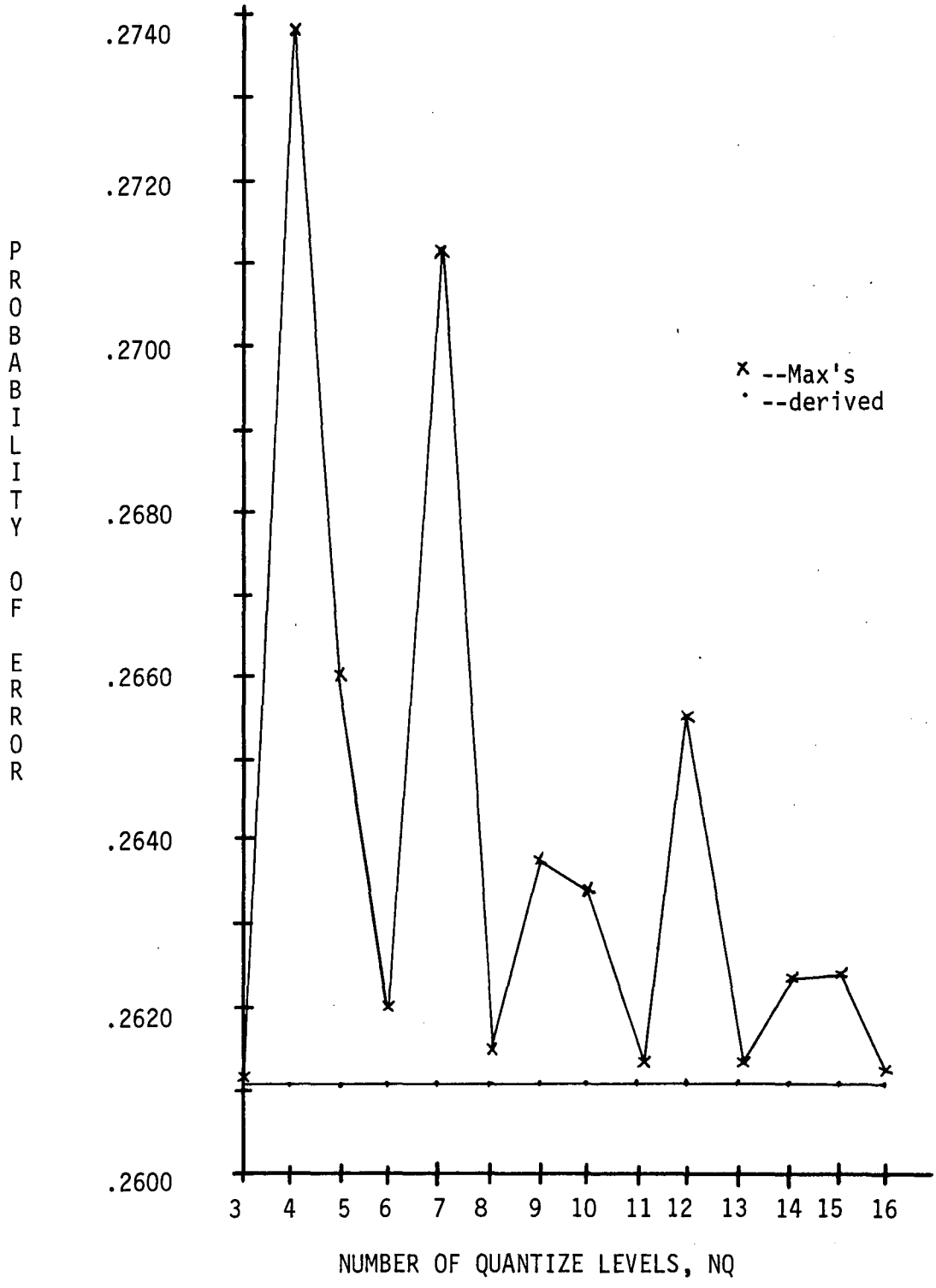

Figure 7.- Probability of error utilizing Max's interval and the derived interval.

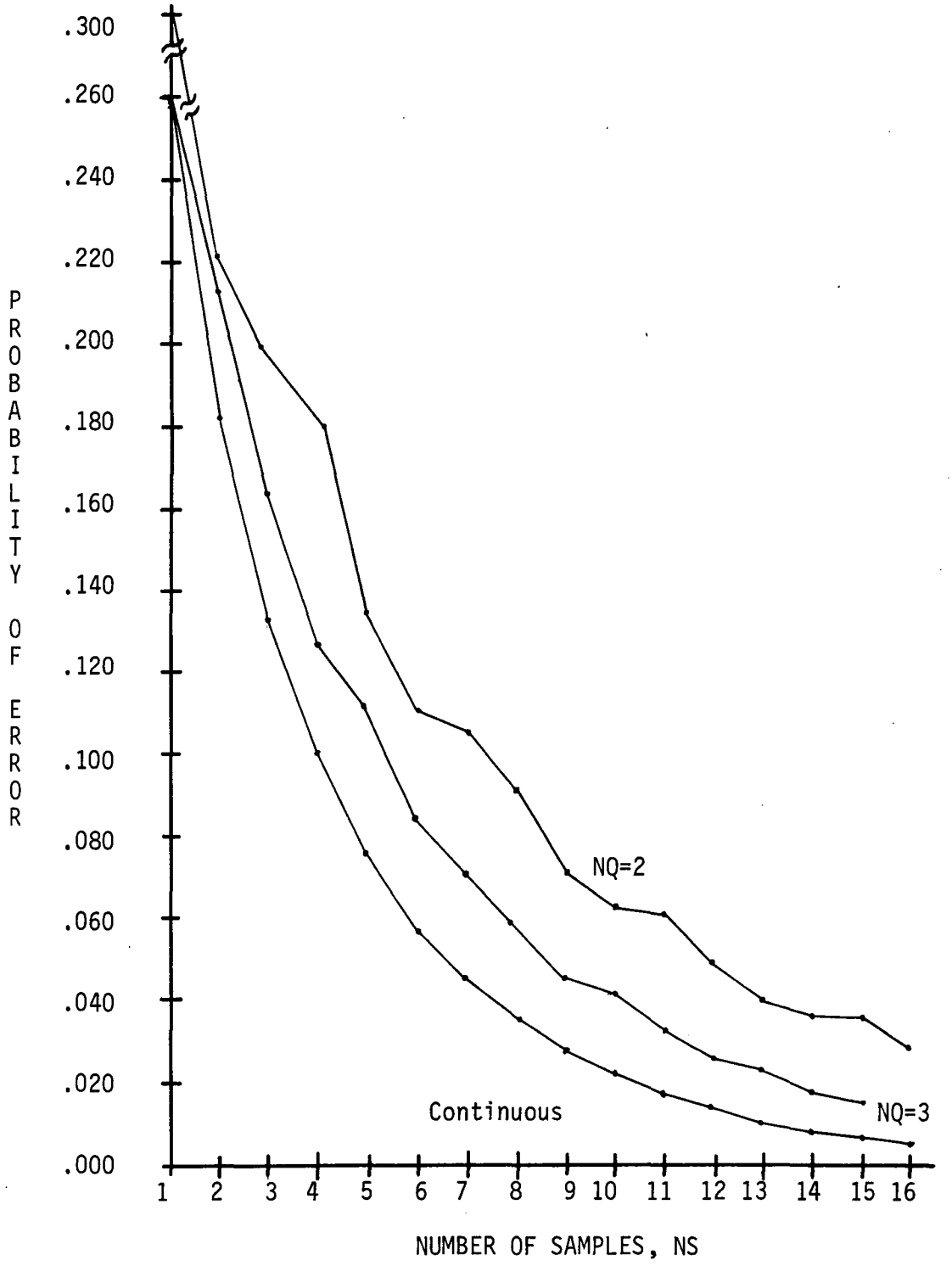

Figure 8.- Probability of error as a function of samples for Max's quantized data and continuous data.

BIBLIOGRAPHY

(1) R. F. Nalepka, "Investigation of Multispectral Discrimination Techniques", Infrared and Optics Laboratory, Willow Run Laboratories, Institute of Science and Technology, Contract #12-14-100-9548(20), January, 1970.

(2) J. R. Welch and K. G. Salter, "A Context Algorithm for Pattern Recognition and Image Interpretation", IEEE Transactions on Systems, Man and Cybernetics, Vol. S MC-1, pp24-30, January, 1971.

(3) J. E. Boyd, "Classification Error of Gaussian and Transformed Gaussian Variates", Master of Science Thesis, Department of Electrical Engineering, South Dakota State University, Brookings, South Dakota, 1971.

(4) A. Papoulis, Probability, Random Variables and Stochastic Processes, New York, McGraw Hill Book Company, 1965.

(5) G. M. Dilliard, "Generating Random Numbers Having Probability Distributions Occuring in Signal Detection Problems", IEEE Transactions on Information Theory, Vol. IT-13, No. 4, pp 616-617, October, 1967.

(6) B. W. Lindgren, Statistical Theory, New York, Macmillan, 1962.

(7) D. V. Serreyn and G. D. Nelson, "Classification Error Using Quantized Data", Remote Sensing Institute, South Dakota State University, Brookings, South Dakota; Interim Technical Report RSI-71-21, November, 1971.

(8) J. Max, "Quantization for Minimum Distortion", IRE Transactions on Information Theory, Vol. IT-16, No. 1, pp 7-12, March, 1960.

(9) E. G. Johnson, "The Pax II Picture Processing System" from Picture Processing and Psychopictories edited by B. S. Lipkin and A. Rozenfeld, Academic Press: New York, 1970, pp. 427-512.

(10) C. J. Frazee, R. D. Heil, and F. C. Westin, "Remote Sensing for Detection of Soil Limitations in Agricultural Areas", Annual Report, Remote Sensing Institute, South Dakota State University, Brookings, South Dakota, June, 1970.